Data Curation Tasks for Large Language Models (LLMs)

Building Training Corpora for LLMs

Ramaseshan Ramachandran

"Garbage In, Garbage Out"

- Curation ensures that the massive datasets used to train LLMs (often trillions of tokens) are of **high quality** and **relevance**.
- It directly impacts the model's:
  1. Accuracy and Coherence
  2. Safety and Ethics
  3. Generalization (avoiding overfitting)

TASK 1: DATA CLEANING AND PREPROCESSING

- **Goal:** Remove noise and standardize the text format.
- **Key Actions:**
  1. **Boilerplate Removal:** Stripping out non-content text from web scraping (e.g., HTML tags, advertisements, menus).
  2. **Text Normalization:**
     - Unifying Unicode representations.
     - Standardizing quotes, dashes, and punctuation.
     - Consistent casing (if necessary).
  3. **Language Identification:** Filtering out documents that are not in the target language for pre-training.

TASK 2: FILTERING AND DEDUPLICATION

- **Goal:** Ensure data quality and prevent memorization/overfitting.
- **Quality Filtering:** Using criteria or models to discard low-quality documents.
  - Documents with excessive grammatical errors or high symbol-to-word ratios.
  - Content that is too short, repetitive, or appears spam-like.
- **Deduplication:**
  - **Exact Match:** Removing identical copies of documents.
  - **Fuzzy Match:** Using techniques (e.g., MinHash, LSH) to identify and remove near-identical documents to improve generalization.

TASK 3: SAFETY AND ETHICAL ALIGNMENT

- **Goal:** Mitigate risks related to privacy, toxicity, and bias.
- **Privacy Filtering (PII Redaction):**
  - Detecting and removing or replacing Personally Identifiable Information (PII), such as names, addresses, and phone numbers.
- **Toxicity Filtering:**
  - Using toxicity classifiers (e.g., Perspective API) to flag and filter out hate speech, explicit, or harmful content.
- **Bias Mitigation:**
  - Identifying and re-weighting or removing biased samples to promote fairness and diverse representation.

POST-TRAINING CURATION: ALIGNMENT DATA

- **Context:** After base model pre-training, specific, high-quality data is curated for alignment.
- **Supervised Fine-Tuning (SFT) Data:**
  - Curating high-quality **Instruction-Response pairs** (prompt $\rightarrow$ desired output).
- **Reinforcement Learning with Human Feedback (RLHF) Data:**
  - **Preference Data:** Human annotators rank multiple model responses to train a **Reward Model (R_θ)**.
  - The quality of this human-labeled data is paramount for safety and helpfulness alignment.
- **Synthetic Data:** Using powerful LLMs to generate high-quality, diverse, and clean training examples (e.g., for instruction tuning).